Synopsis for Registration as Research Student

for

Doctoral Degree

on

Modified Descriptors for Visual Search using Deep Learning

By

Mr. Vikas Dattatray Patil

Research Scholar

Supervisor

Dr. Manesh B. Kokare

Department of Electronics and Telecommunication Engineering

Shri Guru Gobind Singhji

Institute of Engineering and Technology, Vishnupuri,

Nanded – 431606 (Maharashtra)

September, 2022

Abstract

For the search a variety of visual feature extraction techniques have been used. Due to the computation time required, some good algorithms are not used. The retrieval performance of a content-based image retrieval system critically depends on the feature representation and similarity measures. The goal of the proposed method is to develop an efficient algorithm that can handle the above problem. The Deep Learning Network is used for feature extraction and classification and is an emerging research area due to a large amount of data. Searching for a specific object in an image collection has a variety of applications in location search, scene retrieval, landmark recognition, and product search. Camera-equipped devices are becoming ubiquitous platforms for visual search applications.

Various details about a picture are included in an image descriptor, which is a feature vector. It is a concise representation that is useful for differentiating between images. Low bit rate, mobile, visual search has become the preferred approach for small descriptors. Sending compact descriptors is a secure replacement for the time consuming transmission of image queries, which may greatly reduce network latency and enhance user experience.

Introduction

The design and development of CDVS (Compact Descriptors for Visual Search) addresses the standardization of object instance recognition technology, leveraging on a several compelling functionalities and briefly enabling a number of advantages. Compression is an essential requirement for typical client-server visual search architectures. The CDVS is changing the communication paradigm rather than sending a compressed image to a server, a set of local features are extracted from the query image and compressed into a single compact descriptor on the client side, and the resulting compact descriptor is then sent to the server to initiate a search. The overall size of a set of local uncompressed features (e.g. the well-known SIFT descriptor) extracted from an image can be larger than traditionally compressed JPEG files. CDVS allows to drastic decrease of dimension of the compact visual descriptors, thanks to a scalable and adaptive compression scheme. CDVS supports different sizes of compact descriptor footprint, spanning from a maximum of 16 KB per image, which is the fully performing operating mode, down to 512 Bytes, for extremely constrained bandwidth scenarios**.**

The amount of audio-visual information that is available in digital form is significantly growing because of new communication technologies and the widespread usage of the Internet in our society. Therefore to search for and classify a wide variety of multimedia information, it has been required to create certain systems that enable content description. The descriptors enable rapid and effective searches of the audio-visual information since they have a thorough understanding of the objects and events included in a video, picture, or audio file.

Motivation

The size of user's digital media collections is growing exponentially as millions of images and videos are added to servers every day. Navigating through these petabytes of visual data and organizing them is hard without appropriate visual search tools. Compact Descriptors for Visual Search tools support quick and efficient visual search, enabling access to even the largest image databases.

Camera-equipped mobile devices, such as mobile phones or tablets are becoming platforms of choice for the deployment of visual search, augmented reality, and related applications. Many applications involve the analysis of visual information and the exchange of related metadata with remote servers. The scalability, response time, and usability of a multimedia search system depend critically on the format and volume of the data interchanged. Image descriptors capable of matching parts of the depicted scene often reach a size not much smaller than the actual image. Thus, the compactness of such descriptors has become increasingly important, but an interoperable solution did not exist prior to CDVS

Literature Survey

Recent developments in deep learning have shown that it is possible to train a neural network to learn and compare pictures using local descriptors in an automated manner without necessarily depending on handcrafted techniques. The LDVS (Learnable Descriptors for Visual Search), is a learnable binary local descriptor devised for matching natural images within the MPEG CDVS framework [1]. The MPEG CDVS (Compact Descriptors for Visual Search) standard [2] aims at low- complexity, bitrate-efficient, large-scale image matching, and retrieval over mobile devices. Preliminarily, SIFT descriptors [3-4] are extracted from the image and compressed as binary local descriptors. Each image is represented as a set of compressed descriptors, the number and size depending on the target descriptor bitrate, i.e. the encoding mode. Then, compressed descriptors are combined into a global descriptors. The ensemble of local descriptors and global descriptor extracted from an image forms the image CDVS descriptor. Images can be matched by comparing the relative CDVS descriptors in multiple ways, such as matching pairs of local descriptors comparing Hamming distances followed by a geometric consistency check.

The Current research investigated the potential of deep convolutional architectures to learn global descriptors. In [5] it is proposed to learn a global descriptor a trainable neural network functions which aggregate the local descriptors into a global representation. In [6]the

compression of a global CDVS descriptor is formulated as a resource-constrained optimization problem. The CDVS local and global descriptors are complemented with a global descriptor called Nested Invariance Pooling (NIP) learned using a convolutional architecture [7]. SIFT-like learned descriptors retain desirable properties such as invariance to rotations, illumination, and perspective changes [8]. Interesting results are reported when pairs or triplets of patches are jointly encoded, and a decision network is trained to learn an appropriate distance metric [9-10]. While complete image matching pipelines based on a deep learning framework such as a matching network which can produce very accurate results in less than a second of GPU computation [11] go beyond the scope of the present work, it shows the potential of image matching architectures based on deep learning frameworks.

A considerable amount of learned descriptors such as LIFT [12], MatchNet [13], HardNet [14], and many others use Convolutional Neural Networks (CNNs) to learn discriminant patch representations and in particular train a Siamese network with pairs of (non-)corresponding patches [9]-[11], [13]. However, these designs have not gained momentum in practical applications. Binary descriptors allow for fast comparison via Hamming distance, and deployment in limited-resources applications thanks to their improved rate and storage efficiency. A list of seminal works would include early descriptors such as BRIEF [16], BRISK [17], FREAK [18], and ORB [19]. Learned binary descriptors became ever more relevant thanks to works such as BinBoost [20].

Identified Open Research Problems

As reported by Andrea Migliorati etc. Retrieval rate is 37.07 % using LDVS there is scope to increase image retrieval rate by modifying key point extraction strategies of LDVS as bellow

1. There is limited work observed on use descriptor and detector other than LOG+SIFT
2. The less work is published on applying different deep learning algorithm other than CNN.
3. There is not much of work seen on use of other network than the Siamese Network.
4. One can work on minimizing the Quantization Error

Problem Definition

From the development of multimedia databases, the importance of a huge amount of material is determined by how well it can be viewed, retrieved and relevant knowledge extracted from it while reducing the amount of time spent looking at it. Multimedia information is critical in applications like entertainment, education, e-commerce, health, and aerospace in the digital field. With the rapid growth of the Internet, consumers now have access to a massive amount of multimedia material. As a result, many digital images are generated in daily life, and if they are inspected on a regular basis, there is a wealth of helpful information available to consumers.

Methodology

Searching for a specific object in a large collection of images has been a long-standing research problem, with a wide range of applications in location recognition and product search. Much research has been done on different modules such as image representation, matching, indexing, retrieval, and geometric re-ranking. Figure 1 gives a baseline flowchart for proposed visual search. Local features are extracted from database images. The server searches for relevant images based on the visual similarity between a query and database images, and inverted indexing is often employed to speed up the search process. Local features are proven to be particularly important in visual search. Database images containing the nearest neighbors are efficiently collected based on index files and are accordingly ranked by the similarity score. Finally, the top returned images are re-ranked through geometric consistency checking in which the location of local descriptors is considered as well.

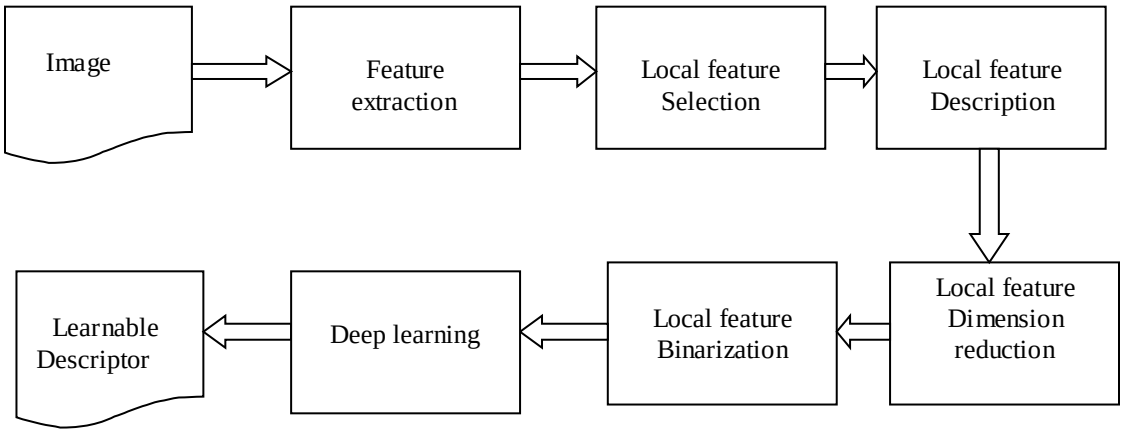

Fig. 1. Block diagram of Deep Learning based MLDVS

A typical MLDVS (Modified Learnable Descriptors for Visual Search) application framework is shown in Figure 1. The local features are extracted and compressed to produce compact visual descriptors on the mobile devices and the retrieval is performed on the remote server using the received descriptors.

Generally, for an image retrieval system, the user supplies a query image or a query object by selecting a region of the query image, and then the feature detection module extracts local features in the from key points.

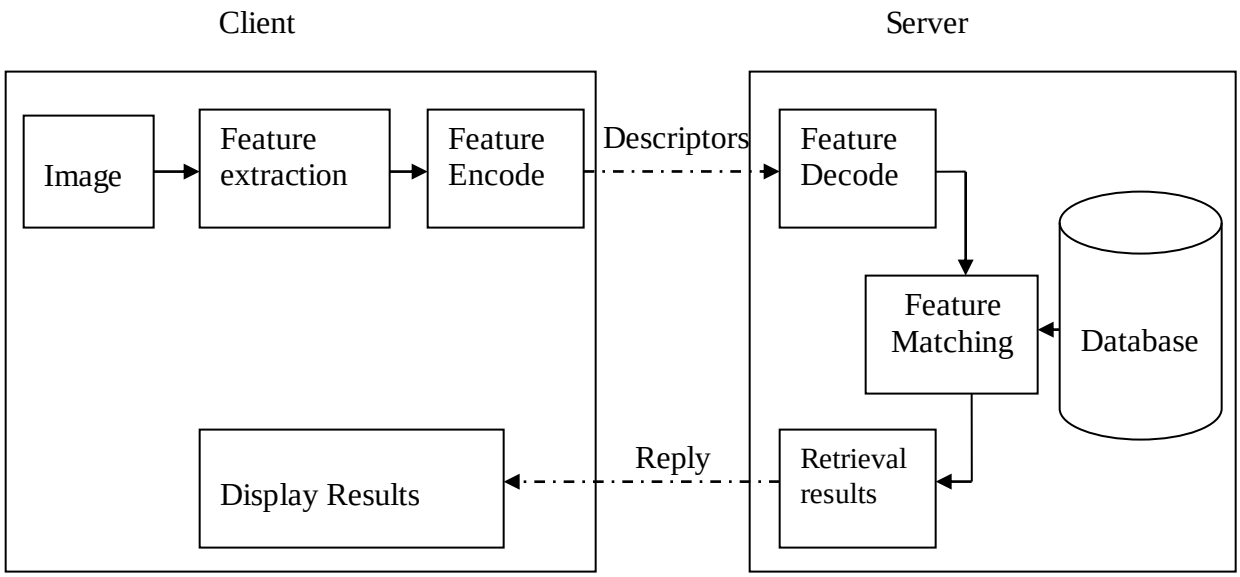

Fig. 2. General Architecture of Proposed System.

Based on modified LVDS as shown in figure 1 it is possible to propose a system as shown in figure 2 in client-server visual search systems, a collection of local features are extracted from the query picture and compressed into a single compact descriptor on the client side. This compact descriptor is then transmitted to the server to begin search instead of sending a compressed image to the server. A set of locally uncompressed characteristics that are taken from an image may have a greater total size than a JPEG file that has been compressed conventionally. It possible to significantly reduce the size of the compact visual descriptors.

Future Research Plan – Objectives

The objectives involved in visual search can be described as follows-

1. To make review and analysis of existing methods for visual search.
2. To propose modified method for descriptor based visual search.
3. To analyse the performance parameters of the modified method.
4. To evaluate the performance of proposed method by comparing it with existing methods.
5. To develop image retrieval system based on descriptor mentioned in objective 2 and using deep learning algorithm.
6. To analysis parameters of designed retrieval system

Conclusion

The proposed MLDVS descriptors are learnable, binary local descriptors designed for image matching within the MPEG CDVS international standard. MLDVS descriptors are learned so that they can be sign-quantized and compared using the Hamming distance as required by the standard. The fully convolutional architecture exhibits a low parameters count with respect to comparable architectures. Then, a complete pair-wise image matching pipeline is designed around the introduced MLDVS descriptors.

Proposed Work Plan

Table 1: Year-wise research work plan

Year	Work
I (2021-22)	Literature Survey, Tools understanding for problem solving, Learning new technologies, Implementation of existing algorithms, Result analysis.
II (2022-23)	working on developing a technique for a targeted problem, comparing results, Writing research articles, and submitting research articles to journals
III (2023-24)	Pre-PhD Synopsis submission, Thesis writing, Thesis submission

References

[1] Andrea Migliorati ,Attilio Fiandrotti ,Gianluca Francini , and Riccardo Leonardi , "Learnable Descriptors for Visual Search" IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 30, 2021 pp.80-91.

[2] G. Francini, M. Balestri, and S. Lepsoy, *CDVS: Telecom Italia Response to CE1-Interest Point Detection*, Standard ISO/IEC JTC1/SC29/WG11 31369, 2013.

[3] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[4] S. Zhang, Q. Tian, K. Lu, Q. Huang, and W. Gao, "Edge-SIFT: Discriminative binary descriptor for scalable partial-duplicate mobile search," *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2889–2902, Jul. 2013.

[5] J. Qian, X. Lin, H. Liu, Y. Deng, and R. Ji, "Towards compact visual descriptor via deep Fisher network with binary embedding," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2018, pp. 1–6.

[6] Y. Wu *et al.*, "Codebook-free compact descriptor for scalable visual search," *IEEE Trans. Multimedia*, vol. 21, no. 2, pp. 388–401, Feb. 2019.

[8] Y. Lou *et al.*, "Compact deep invariant descriptors for video retrieval," in *Proc. Data Compress. Conf. (DCC)*, Apr. 2017, pp. 420–429.

[9] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, and F. Moreno-Noguer, "Discriminative learning of deep convolutional feature point descriptors," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*,Dec. 2015, pp. 118–126.

[10] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis.Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4353–4361.

[11] W. Luo, A. G. Schwing, and R. Urtasun, "Efficient deep learning for stereo matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.(CVPR)*, Jun. 2016, pp. 5695–5703.

[12] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua, "LIFT: Learned invariant feature transform," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 467–483.

[13] X. Han, T. Leung, Y. Jia, R. Sukthankar, and A. C. Berg, "MatchNet: Unifying feature and metric learning for patch-based matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015,pp. 3279–3286.

[14] A. Mishchuk, D. Mishkin, F. Radenovic, and J. Matas, “Working hard toknow your neighbor’s margins: Local descriptor learning loss,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 4829–4840.

[15] J. Žbontar and Y. LeCun, “Stereo matching by training a convolutional neural network to compare image patches,” *J. Mach. Learn. Res.*, vol. 17, nos. 1–32, p. 2, 2016.

[16] M. Calonder, V. Lepetit, M. Ozuysal, T. Trzcinski, C. Strecha, and P. Fua,“BRIEF: Computing a local binary descriptor very fast,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1281–1298, Jul. 2012.

[17] S. Leutenegger, M. Chli, and R. Y. Siegwart, “BRISK: Binary robust invariant scalable keypoints,” in *Proc. Int. Conf. Comput. Vis.*,Nov. 2011, pp. 2548–2555.

[18] A. Alahi, R. Ortiz, and P. Vandergheynst, “FREAK: Fast retina keypoint,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012,pp. 510–517.

[19] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, “ORB: An efficient alternative to SIFT or SURF,” in *Proc. Int. Conf. Comput. Vis.*,Nov. 2011, pp. 2564–2571.

[20] T. Trzcinski, M. Christoudias, P. Fua, and V. Lepetit, “Boosting binary keypoint descriptors,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2874–2881.

Mr. Vikas Dattatray Patil
(Research Scholar)

Dr. Manesh B. Kokare
(Supervisor)